Basic Statistics

732A47 Text Mining

Overview

Probability theory
- Probability
- Distributions and their properties
- Bayes theorem
- Generating test data

Statistical Inference
- Hypothesis testing
- Linear regression
- Logistic regression
- Support vector machines

Probability

How likely it is that some event will happen?

Idea:
- Experiment
- Outcomes (sample points) $O_1, O_2, \ldots O_n$
- Sample space Ω
- Event A
- Probability function P: Events → [0,1]

Example: Tossing the coin

Properties and definitions

- One can think of events as sets
  - Set operations are defined: $A \cup B, A \cap B, \bar{A} \backslash B$
- $P(A \cup B) = P(A) + P(B)$ if $A \cap B = \emptyset$
- Independence $P(A, B) \equiv P(A \cap B) = P(A)P(B)$
- Conditional probability $P(A|B) = \frac{P(A,B)}{P(B)}$

Bayes theorem

Example:
- We have constructed spam filter that
  - identifies spam mail as spam with probability 0.95
  - Identifies usual mail as spam with probabilitt 0.005
- This kind of spam occurs once in 100,000 mails
- If we found that a letter is a spam, what is the probability that it is actually a spam?

Bayes theorem

- We have some knowledge about event B
  - Prior probability P(B) of B
- We get new information A
  - P(A)
  - P(A|B) probability of A can occur given B has occured
- New (updated) knowledge about B
  - Posterior probability P(B|A)

$$P(B|A) = \frac{P(A|B)P(B)}{P(A)}$$

Bayes theorem for events

- If B_i are disjoint and $A \subseteq \cup B_i$ then

$$P(B_j|A) = \frac{P(A|B_j)P(B_j)}{\sum_i P(A|B_i)P(B_i)}$$

Note, $\cup B_i$ can be equal to Ω

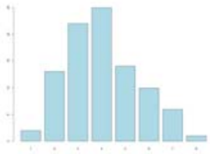
Random variables

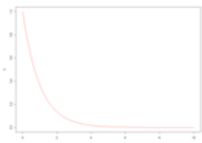
- Instead of having events, we can have a variable X:
  - Events→$\mathbb{R}$ Continuous random variables
  - Events→$\mathbb{N}$ Discrete random variables

Examples:
- X={amount of times the word "crisis" can be found in financial documents}
  - P(X=3)
- X={Time to download a specific file to a specific computer}
  - P(X=0.36 min)

Distributions

- Discrete
  - Probability mass function P(x) for all feasible x
- Coninuous
  - Probability density function f(x)
  - Cumulative density function $F(x) = \int_0^x f(t)dt$





Expected value and variance

- Expected value = mean value
  - $E(X) = \sum_{i=1}^{n} X_i P(X_i)$
  - $E(X) = \int X f(X) dX$
- Variance how much values of random variable can deviate from mean value
  - $Var(X) = E\big(X - E(X)\big)^2 = E(X^2) - E(X)^2$

Some conventional distributions

Bernoulli distribution

- Events: Success (X=1) and Failure(X=0)
- P(X=1)=p, P(X=0)=1-p
- $E(X) = p$
- $Var(X) = 1 - p$

Examples: Tossing coin, vinning a lottery,..

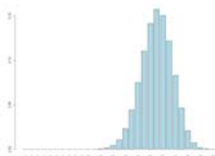
Some conventional distributions

Binomial distribution

- Sequence of *n* Bernoulli events
- X={Amount of successes among these events}, X=0,…,n

$$P(X = r) = \frac{n!}{(n-r)!\,r!} p^r (1-p)^{n-r}$$

- $EX = np$
- $Var(X) = np(1 - p)$



Poisson distribution

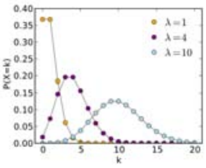
- Customers of a bank **n** (in theory, endless population)
- Probability that a specific person will make a call to the bank between 13.00 and 14.00 a certain day is ***p***
  - ***p*** can be very small if population is large (rare event)
  - Still, some people will make calls between 13.00 and 14.00 that day, and their amount may be quite big
  - A known quantity **λ=*np*** is mean amount of persons that call between 13.00 and 14.00
  - **X**={amount of persons that have called between 13.00 and 14.00}

Poisson distribution

- $P(X=r) = \lim_{n\to\infty} \frac{n!}{(n-r)!r!} p^r (1-p)^{n-r}$
- It can be shown that

$$P(X=r) = \frac{\lambda^r e^{-\lambda}}{r!}$$

- $E(X) = \lambda$
- $Var(X) = \lambda$



Poisson distribution

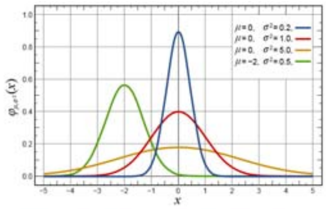
- Further properties:
  - Poisson distribution is a good approximation of the binomial distribution if n >20 and p < 0.05
  - Excellent approximation if n ≥ 100 and np ≤ 10

Normal distribution

- Appears in almost all applications
  - Time required to download a specific document to a specific computer

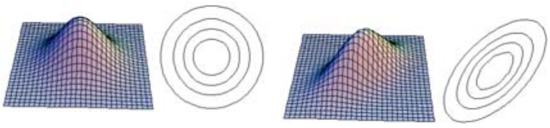
$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \sigma > 0$$

- $E(X) = \mu$
- $Var(X) = \sigma^2$



Multivariate distributions

- Probability of two variables having certain values at the same time
  - P.D.F. f(x,y)
  - Correlation



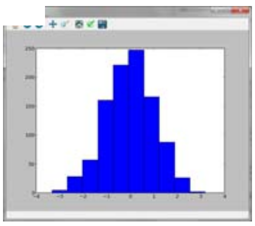
Radnom number generation

- Random number (sequences) can be generated in computer but those are not truly random!
  - Deterministic algorithms are used
  - Current time is also used to mimic randomness

```
>>> from scipy.stats import norm
>>> norm.pdf([-1., 0, 1])
array([ 0.24197072,  0.39894228,  0.24197072])
>>> norm.pdf(0, loc=1, scale=2)
0.17603266338214976
>>> t=norm.rvs(loc=2,scale=0.1,size=10)
>>> t
array([ 2.12425086,  1.81937297,  2.15002352,  2.086603  ,  2.1119639 ,
        1.8902901 ,  1.98702357,  2.03978701,  2.00005928,  2.13550148])
>>>
```


Generating a sample

```
>>> from scipy.stats import norm
>>> t=norm.rvs(size=1000)
>>> hist(t)
```



Generating datasets

- Necessary for testing different algorithms
- Dataset $D = \{(X_i, Y_i), i = 1, \ldots, n\}$

Case	Variable 1	Variable 2	Variable 3	.	.	.	
1							
2							
3							
.							
.							
.							

- Decide
  - the structure of each input variable
  - the distribution of some input variables
  - multivariate distribution for some subset of input variables
  - Response model

Generating data

- Response models
  - Continuous variables
    - $Y = f(\boldsymbol{X})$ or $y = f(\boldsymbol{X}) + \epsilon$
  - Discrete variables
    - $Y = f(\boldsymbol{X})$ or $y = Generator_P(f(\boldsymbol{X}))$

Hypothesis testing

- Certain proportion of population satisfies some property in our data, *p*
- The true distribution is $Bernoulli(\pi)$
- We have assumption about π:
  - Test $\pi = \pi_0$ vs $\pi \neq \pi_0$
- Special statistical procedures exit to find it out
  - Compute $s = \sqrt{p(1-p)}$
  - Compute the absolute value of $z = \frac{p-\pi_0}{\sqrt{\frac{s^2}{n}}}$ and compare it with value from a table
  - If this value is greater the table value, reject $\mu = 0.5$

Chi-square test

H0: Frequencies of words are approximately same in all documents

Ha: Some documents have a different pattern

Word	Document1	Document 2
business	43	66
school	38	72
dollar	11	23
market	8	19

$$\chi^2 = \sum \frac{(O-E)^2}{E}$$

Simple linear regression

Model:

$$y = \beta_0 + \beta_1 x + error$$

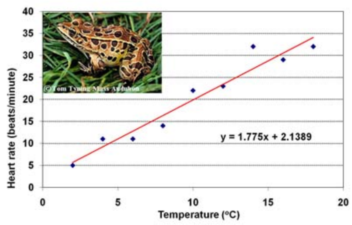
Terminology:

β_0: intercept (or bias)

β_1: regression coefficient

Response

The target responds directly and linearly to changes in the input

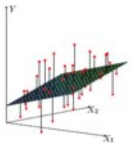


Ordinary least squares regression (OLS)

Model:

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p + error$$

$$y = \beta_0 + \boldsymbol{\beta}^T \boldsymbol{X} + error$$



Find coefficients by minimizing RSS:

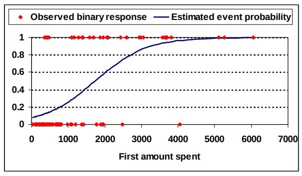
$$RSS(\boldsymbol{\beta}) = \sum_{i=1}^{N} (y_i - \beta_0 - \beta_j \sum_{j=1}^{p} x_{ij})^2$$

Logistic regression: two classes

- Consider Logistic model with predictor X=First_Amount_Spent
- Logistic model

$$\log \frac{P(Y=1 \mid X=x)}{P(Y=0 \mid X=x)} = \log \frac{P(Y=1 \mid X=x)}{1-P(Y=0 \mid X=x)} = \beta_0 + \beta_1 x$$

$$P(Y=1 \mid X=x) = \frac{\exp(\beta_0 + \beta_1 x)}{1+\exp(\beta_0 + \beta_1 x)}$$

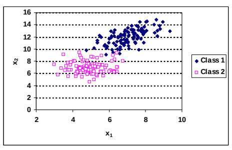


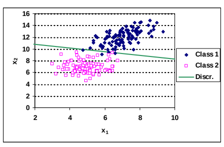
Classification

Given a dataset $D = \{(X_i, Y_i), i = 1, \ldots, n\}$

- X1=frequency of "money"
- X2=frequency of "crisis"
- Y=0 if document is from financial area
- Y=1 if document is from "psychology"

Aim: divide input space into areas in order to predict new observations



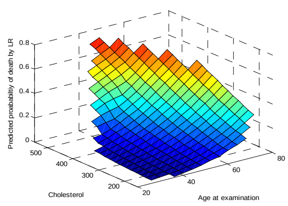


Logistic regression: two classes

- Two predictors
  - Fitted probability of death in coronary heart disease as a function of cholesterol level and age at examination

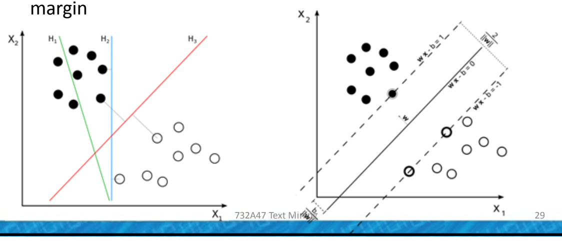
$$\log \frac{P(Y=1 \mid X=\mathbf{x})}{P(Y=0 \mid X=\mathbf{x})} = \beta_0 + \beta_1^T \mathbf{x}$$

$$P(Y=1 \mid X=\mathbf{x}) = \frac{\exp(\beta_0 + \beta_1^T \mathbf{x})}{1+\exp(\beta_0 + \beta_1^T \mathbf{x})}$$



Support vector machines

- Very good in classification of high dimensional data→ very popular for text classification
- Support vector=observations on the boundary of classes
- **Linear SVM model**: find a linear boundary having maximal margin



Linear SVM

- Boundary is given by equation:
  - $\boldsymbol{w} \cdot \boldsymbol{x} + b \ = 0$
- Condition for Y=1:
  - $\boldsymbol{w} \cdot \boldsymbol{x} + b \geq 1$
- Condition for Y=-1:
  - $\boldsymbol{w} \cdot \boldsymbol{x} + b \leq -1$

$$\min_{w} \|w\|^2$$
$$subject\ to\ y_i(\boldsymbol{w}\boldsymbol{x}_i + \boldsymbol{b}) \geq 1, \qquad i = 1, \ldots, n$$

→Convex optimization problem

We need to find *w* and *b* that have largest margin $\frac{2}{\|\boldsymbol{w}\|}$

SVM

- Compared to many other methods, a global minimum of the objective function is possible to find
- There are generalizations of SVM for
  - Nonlinear boundaries
  - More than two classes
  - Cases when the boundaries are not separable

Reading

- www.scipy.org
- MS, Chapter 2, 5.1-5.3